(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－244693

(43)公開日　平成９年(1997) ９月19日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ１０Ｌ　5/04			Ｇ１０Ｌ　5/04	Ｄ	
				Ｆ	

審査請求　未請求　請求項の数６　ＯＬ　（全 ７ 頁）

(21)出願番号　特願平8－49774

(22)出願日　平成８年(1996) ３月７日

(71)出願人　000102728
エヌ・ティ・ティ・データ通信株式会社
東京都江東区豊洲三丁目３番３号

(72)発明者　新村　貴彦
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内

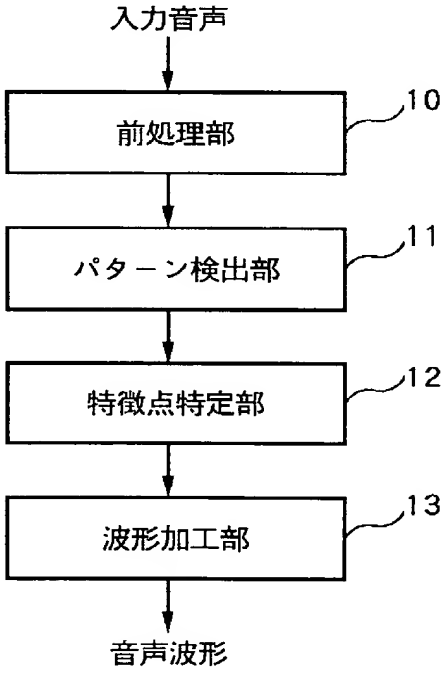
(74)代理人　弁理士　鈴木　正剛

(54)【発明の名称】　音声合成方法及び装置

(57)【要約】

【課題】　音声合成装置において、音声モーフィング等の音声変換を行う際の合成音声の明瞭性を向上させる。

【解決手段】　声質の異なる二種以上の入力音声をそれぞれ音韻毎に区切り、対応する音韻毎に各入力音声を合成することで新たな音声を得る。その際、前処理部１０で、文章を音韻単位に区切り、各音韻単位に属するピッチパターンをパターン検出部１１で求める。特徴点特定部１２では、ピッチパターンの特徴点、例えば極値のピッチ波形を入力音声毎に対応して特定する。波形加工部１３では、対応するピッチ波形毎に合成処理を行って各音韻の合成処理を行う。



【特許請求の範囲】
【請求項１】　声質の異なる二種の入力音声をそれぞれ所定区間単位で区切り、個々の区間単位に対応するピッチ波形を組み合わせて合成音を生成する方法であって、個々の入力音声に属する前記区間単位内のピッチ波形のパターン（以下、ピッチパターン）の変化傾向を検出する過程と、
検出したピッチパターンの変化傾向の特徴点を区間単位毎に特定するとともに、各入力音声毎に、それぞれ特定された特徴点のピッチ波形及び特徴点間のピッチ波形に基づく新たなピッチパターンを生成する過程と、
を含むことを特徴とする音声合成方法。
【請求項２】　前記生成された新たなピッチパターンに対応する元のピッチパターンに所定係数を乗じて新たなピッチ波形を生成する過程を含むことを特徴とする請求項１記載の音声合成方法。
【請求項３】　互いに声質の異なる第一及び第二の入力音声に対し、第一の入力音声から第二の入力音声へのモーフィング処理を行う方法であって、
前記第一及び第二の入力音声をそれぞれ所定区間単位で区切り、個々の区間単位内のピッチパターンの変化傾向を検出する過程と、
検出したピッチパターンの変化傾向の特徴点を区間単位毎に特定するとともに、各入力音声毎に、それぞれ特定された特徴点のピッチ波形及び特徴点間のピッチ波形に基づく新たなピッチパターンを生成する過程と、
生成された新たなピッチパターンに対応する元のピッチパターンに所定係数を乗じてモーフィング処理に用いるピッチ波形を生成する過程とを含み、
前記第二の入力音声に対する前記係数をモーフィングが進むにつれて大きくすることを特徴とする音声合成方法。
【請求項４】　前記特徴点が個々の区間単位におけるピッチパターンの極値であることを特徴とする請求項１ないし３のいずれかの項記載の音声合成方法。
【請求項５】　前記新たなピッチパターンにおける特徴点間のピッチ波形を、前記入力音声のピッチ波形の数をもとに補間することを特徴とする請求項１ないし４のいずれかの項記載の音声合成方法。
【請求項６】　声質の異なる二種の入力音声をそれぞれ所定区間単位で区切る前処理部と、
この前処理部で区切られた個々の区間単位内のピッチパターンの変化傾向を検出するパターン検出部と、
このパターン検出部で検出したピッチパターンの変化傾向の特徴点を区間単位毎に特定する特徴点特定部と、
この特徴点特定部により各入力音声毎に特定された特徴点のピッチ波形及び特徴点間のピッチ波形に基づく新たなピッチ波形を生成する波形加工部と、
を有し、この新たなピッチ波形を組み合わせて合成音を生成することを特徴とする音声合成装置。

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、音声合成技術に関し、例えば、ある人物の音声を別人の音声へと変えていく音声モーフィング技術に関する。
【０００２】
【従来の技術】音声合成技術は、駅構内でのアナウンスや機械による文章朗読等に広く用いられている。近年は、マルチメディア技術の台頭により、音声の表現力をより高めることが要求されてきており、例えば音声モーフィング処理等に代表される、声質変換という高度な技術が求められている。
【０００３】音声モーフィング処理とは、画像処理技術において、ある人物の顔画像を徐々に別人の画像に変えていく過程を表現する画像モーフィング技術を音声に適用したもので、ある人物の声質を別人の声質へと変えていく過程を表現するものである。以下、音声モーフィングという場合、一の人間の声質を他の人間の声質に徐々に変えていくことをいうものとする。従来より、このような音声モーフィングは少なからず提案されており、例えば入力音声を音韻毎に区切って時間軸上で音声のスペクトグラム成分を線形に変化させるものが知られている。この技術において、読み上げ対象となる文を、最初は人物Ａの音声により読み上げ、文の後半部は人物Ｂの音声に声質を変える場合の処理を以下に示す。
【０００４】まず、人物Ａ，Ｂのそれぞれに予め読み上げ対象の文を読み上げてもらい、そのピッチ波形を格納しておく。そして、読み上げ文の前半部は人物Ａ、後半部は人物Ｂのピッチ波形をそのまま用い、 中間部には、人物Ａ，Ｂのピッチ波形を合成して得られる合成音声を用いて音声モーフィング処理を行う。
【０００５】この処理の具体例を図８及び図９を用いて説明する。図８では「東京地方は多少雲が多いものの晴れ間も出ています」という文の読み上げの際に、音声モーフィング処理を行う場合の例を示すものである。この図に示されるように、文の前半部は人物Ａの音声をそのまま用い、変換部分となる中間部、即ち「多い（おおい）」にあたる部分は合成音声により読み上げ、後半部を人物Ｂの音声をそのまま用いている。
【０００６】図９に示されるように、上記音声の合成処理においては、まず人物の音韻（/o/，/o/,/i/）毎に音声を区切り、それぞれ実線矢印で示されるように、ピッチ波形の対応をとる。そして、対応するピッチ波形を足し合わせて、破線矢印で示される合成ピッチ波形を得る。次に、各音韻毎に、単純にピッチ波形の周波数の平均値を計算して点ピッチ周波数を求める。そして、人物Ａ，Ｂの点ピッチ周波数の平均値を合成音のピッチ波形の周波数として設定する。さらに、合成したピッチ波形を、もとの語順に従って接続して合成音声を得る。
【０００７】

【発明が解決しようとする課題】上述した従来の音声モーフィング処理においては、各音韻毎に合成音声のピッチ波形の周波数を算出している。この周波数は、ピッチ波形の間隔（時間長）の逆数となる。具体的には、図１０（ａ），（ｂ）に示されるように、人物Ａ，Ｂの音韻のピッチ波形毎に、図中の丸印で示される周波数をそれぞれ求め、さらに、図中の点線で示されるように、それぞれ点ピッチ周波数を算出している。合成音声のピッチ波形毎の周波数は、図１０（ｃ）に示されるように、人物Ａ，Ｂの点ピッチ周波数の平均値を用い、各音韻について固定値としている。しかし、従来の音声モーフィング処理では、音韻をつなげて文章読み上げを行った場合に、その明瞭性が損なわれてしまうという難点が生じていた。

【０００８】本発明の課題は、例えば音声モーフィング等の声質変換を行う際の合成音声の明瞭性を向上させる技術を提供することにある。

【０００９】

【課題を解決するための手段】本発明者らによる検証の結果、従来の方法で明瞭性が劣るのは、音韻毎の点ピッチ周波数が固定値であり、点ピッチ周波数の変化が単調となる点に起因することが判明した。そこで、本発明では、音韻内での点ピッチ周波数を変化させる、改良された音声合成方法及び装置を創案した。

【００１０】即ち、本発明の音声合成方法では、声質の異なる二種の入力音声をそれぞれ所定区間単位で区切り、個々の区間単位に対応するピッチ波形を組み合わせて合成音を生成する。その際、個々の入力音声に属する前記区間単位内のピッチパターンの変化傾向を検出する過程と、検出したピッチパターンの変化傾向の特徴点を区間単位毎に特定するとともに、各入力音声毎に、それぞれ特定された特徴点のピッチ波形及び特徴点間のピッチ波形に基づく新たなピッチパターンを生成する過程と、を含むことを特徴とする。

【００１１】上記方法の後続処理としては、前記生成された新たなピッチパターンに対応する元のピッチパターンに所定係数を乗じて新たなピッチ波形を生成する過程が挙げられる。

【００１２】本発明の他の音声合成方法は、互いに声質の異なる第一及び第二の入力音声に対し、第一の入力音声から第二の入力音声へのモーフィング処理を行う方法であって、前記第一及び第二の入力音声をそれぞれ所定区間単位で区切り、個々の区間単位内のピッチパターンの変化傾向を検出する過程と、検出したピッチパターンの変化傾向の特徴点を区間単位毎に特定するとともに、各入力音声毎に、それぞれ特定された特徴点のピッチ波形及び特徴点間のピッチ波形に基づく新たなピッチパターンを生成する過程と、生成された新たなピッチパターンに対応する元のピッチパターンに所定係数を乗じてモーフィング処理に用いるピッチ波形を生成する過程とを含み、前記第二の入力音声に対する前記係数をモーフィングが進むにつれて大きくすることを特徴とする。

【００１３】上記各方法の好ましい態様としては、前記特徴点を個々の区間単位におけるピッチパターンの極値とする。また、前記新たなピッチパターンにおける特徴点間のピッチ波形を、前記入力音声のピッチ波形の数をもとに補間する。

【００１４】このようにして生成される音声は、元の入力音声のピッチパターンを反映して変化するものとなり、合成音の明瞭性が従来よりも向上する。特に、特徴点として、個々の区間単位におけるピッチ周波数の極値を用いることで、もとの音声の特徴が一層反映された音声が生成される。さらに、新たなピッチパターンにおける特徴点間のピッチ波形を補間することで、元の入力音声の特徴を残しつつ、新たなピッチ波形が生成可能となる。

【００１５】本発明は、また、上記各方法を実施する上で好適となる音声合成装置をも提供する。この装置は、声質の異なる二種の入力音声をそれぞれ所定区間単位で区切る前処理部と、この前処理部で区切られた個々の区間単位内のピッチパターンの変化傾向を検出するパターン検出部と、このパターン検出部で検出したピッチパターンの変化傾向の特徴点を区間単位毎に特定する特徴点特定部と、この特徴点特定部により各入力音声毎に特定された特徴点のピッチ波形及び特徴点間のピッチ波形に基づく新たなピッチ波形を生成する波形加工部と、を有し、この新たなピッチ波形を組み合わせて合成音を生成することを特徴とする。

【００１６】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を詳細に説明する。この実施形態では、従来例と同様、「東京地方は多少雲が多いものの晴れ間もでています」という文章を例として説明する。

【００１７】予め、この文章を人物Ａ，Ｂにそれぞれ読み上げてもらい、その音声を入力音声としてそれぞれ格納しておく。この文章の前半の読み上げには、人物Ａの音声をそのまま用い、後半の読み上げには、人物Ｂの音声をそのまま用いる。一方、この文章の中間部である「多い」の読み上げには、人物Ａ，Ｂの音声から生成される音声を用い、人物Ａから人物Ｂへの声質の変換を行う。その際、中間部の音声は、図１に示す構成の音声合成装置により合成した。

【００１８】この音声合成装置は、入力音声の波形を所定区間単位、例えば音韻を区間単位として区切る前処理部１０と、個々の区間単位（音韻）内のピッチパターンの変化傾向を検出するパターン検出部１１、検出したピッチパターンの変化傾向の特徴点を区間単位毎に特定する特徴点特定部１２と、各入力音声毎に特定されたた特徴点のピッチ波形及び特徴点間のピッチ波形に基づく新たなピッチパターンを生成するとともに、新たなピッチ

パターンに対応する元のピッチパターンに所定係数、例えばある倍率数値を乗じて新たなピッチ波形を生成する波形加工部１３とを備えている。

【００１９】この実施形態では、図２（ａ），（ｃ）に示される人物Ａ，Ｂの音韻に基づいて同（ｂ）に示される合成音の音韻を得るため、上記ピッチパターンの変化傾向を特定するための特徴点として、ピッチパターンの極値を用いた。また、合成されたピッチ波形の周波数として、もとの各ピッチ波形の周波数の平均値を用いた。

【００２０】以下、本実施形態による処理を、図３～図６をも参照して説明する。図３は、音声合成装置の全体的な処理概要を示すフローチャートであり、まず、前処理部１０において、人物Ａ，Ｂそれぞれの音声波形のうち、「多い」に相当する部分を音韻毎に区切る（Ｓ１０１）。音韻毎に区切られた音声波形を図４に示す。図４（ａ）は人物Ａの音声波形、同（ｂ）は人物Ｂの音声波形である。次に、パターン検出部１１において、各音声波形から、ピッチ波形毎の周波数を求める（Ｓ１０２）。これはピッチ波形の間隔から容易に算出することができる。

【００２１】その後、特徴点特定部１２において、音韻の中で周波数が極値をとるピッチ波形を、人物Ａ，Ｂの各音声波形毎に対応をとって特定する（Ｓ１０３、Ｓ１０４）。図５はこの様子を示すもので、（ａ），（ｂ）は人物Ａの音韻のピッチパターン及びそのピッチ波形、（ｄ），（ｃ）は人物Ｂの音韻のピッチパターン及びそのピッチ波形である。この例では人物Ａのピッチ波形１と人物Ｂのピッチ波形１、人物Ａのピッチ波形３と人物Ｂのピッチ波形５、及び人物Ａのピッチ波形７と人物Ｂのピッチ波形９とがそれぞれ対応している。なお、他のピッチ波形に関しては、後述するように、もとの時系列の順に、適宜各ピッチ波形同士を対応させる（Ｓ１０５）。

【００２２】次に、対応するピッチ波形を加算し、もとのピッチ波形の周波数の平均値を求めて合成ピッチ波形の周波数を決定する（Ｓ１０６）。この実施形態においては、図６に示されるように、各ピッチ波形に窓関数をかけたうえで加算することで、合成音声のピッチ波形を得た。また、振幅を正規化するため、振幅幅には係数である”０．５”を乗算した。

【００２３】波形加工部１３では、公知のピッチ波形重畳法を用いて、上述のようにして得られた合成ピッチ波形を、その周波数の間隔で並べ、図７に示される合成音を得る。このようにして、１つ１つのピッチ波形に対して、それぞれ独立に周波数が与えられる。

【００２４】以上のように、本実施形態の音声合成装置では、音韻を構成するピッチ波形毎にそれぞれ周波数が与えられているので、音韻内での周波数が従来のように一定となってしまうことはない。特に、この音声合成装置においては、周波数が極値をとるピッチ波形同士を対応させるようにしているので、各ピッチ波形における周波数が平均化されてしまうこともなく、周波数の高低差を損なうことなく合成音声を得ることができる。

【００２５】なお、この音声合成装置においては、ピッチパターンが極値となるピッチ波形以外の各波形は、以下のように対応させることで補間した。まず、人物Ａ，Ｂの音韻のピッチ波形毎に周波数を計算し、ピッチパターンの極値を求める。図５（ａ），（ｄ）に示されるように、この例においては、人物Ａのピッチ波形は、１、３、７本目で極値をとり、人物Ｂのピッチ波形は、１、５、９本目で極値をとる。これら各ピッチ波形を、極値をとるピッチ波形によって、人物Ａのピッチ波形は３本と４本に、人物Ｂの場合は５本と４本に、それぞれグループ分けする。

【００２６】この際、人物Ａの３本のピッチ波形と人物Ｂの５本のピッチ波形とを対応させるには、波形数の多いほうを分母、少ないほうを分子にして分数を決め、”１”から”５”までをかける。端数は切り捨て、”１”未満は”１”にする。これにより、人物Ａ，Ｂによるピッチ波形の対応をとることができる。同様にして、人物Ａの４本のピッチ波形と人物Ｂの４本のピッチ波形とを対応させる。以下、その具体的対応を示す。

【００２７】まず、人物Ｂの１本目のピッチ波形は、１×３／５＝０．６なので人物Ａの１本目のピッチ波形に対応させる。同様に、人物Ｂの２本目のピッチ波形は２×３／５＝１．２なので人物Ａの１本目、人物Ｂの３本目のピッチ波形は３×３／５＝１．８なので人物Ａの１本目、人物Ｂの４本目のピッチ波形は４×３／５＝２．４なので人物Ａの２本目、人物Ｂの５本目のピッチ波形は５×３／５＝３．０なので人物Ａの３本目、にそれぞれ対応させる。

【００２８】次に、人物Ｂの６本目（Ｂの第二グループの一本目）のピッチ波形は１×４／４＝１なので人物Ａの４本目（Ａの第二グループの一本目）に対応させる。人物Ｂの７本目のピッチ波形は２×４／４＝２なので人物Ａの５本目、人物Ｂの８本目のピッチ波形は３×４／４＝３なので人物Ａの６本目、人物Ｂの９本目のピッチ波形は４×４／４＝４なので人物Ａの７本目にそれぞれ対応させる。こうして、極値の対応を残したままで人物Ａの音声と人物Ｂの音声との対応をとることができる。

【００２９】なお、このような各波形の対応のさせ方については、特に限定がない。例えば、上記例では端数を切り捨てとしたが、端数を切り上げとしてもよいのはいうまでもない。また、この例では人物Ａの声質から人物Ｂの声質への変形ステップが一ステップとなっているが、複数ステップにより声質を変形させてもよい。また、この実施形態では、もとのピッチ波形の点ピッチ周波数の単純平均値を合成音における周波数としているが、この周波数として、人物Ａ，Ｂの点ピッチ周波数の加重平均値を用いてもよい。例えば人物Ｂの入力音声に

対する上記係数をモーフィングが進むにつれて大きくなるようにしてもよい。このようにして、合成音の前半は人物Ａの周波数の影響を強くし、後半は人物Ｂの周波数の影響を強くすることで、人物Ａから人物Ｂへの音声モーフィングをより一層滑らかに行うことも可能である。

【００３０】

【発明の効果】以上の説明から明らかなように、本発明によれば、音韻内での点ピッチ周波数が一定となってしまうことはなく、従って、音韻の明瞭性が従来よりも向上する効果がある。特に、ピッチパターンの極値となるピッチ波形同士を対応させて波形加工を行うようにしたので、周波数の高いもの同士、低いもの同士でピッチ波形の合成がなされる。従って、周波数が高いピッチ波形と低いピッチ波形とが合成されることによる周波数の平均化現象が起こることもなくなり、音声の明瞭性が一層向上する。

【００３１】また、音韻の周波数変化を利用して音声合成を行っているので、顔の動画像と組み合わせたときに、話者の唇の動きと細かな同期を取りやすくなる効果もある。

【図面の簡単な説明】

【図１】本発明の一実施形態の音声合成装置の要部ブロック構成図。

【図２】（ａ），（ｃ）は人物Ａ，Ｂの音声についての音韻のピッチパターンを示すグラフ、（ｂ）は合成音についての音韻のピッチパターンを示すグラフ。

【図３】本実施形態による音声合成装置の処理概要を示すフローチャート。

【図４】（ａ）は合成対象となる音韻についての人物Ａの音声波形、（ｂ）は人物Ｂの音声波形を示す説明図。

【図５】（ａ），（ｂ）は人物Ａについての音韻のピッチパターンとピッチ波形、（ｃ），（ｄ）は人物Ｂについての音韻のピッチパターンとピッチ波形を示す説明図。

【図６】合成音のピッチ波形を得るまでの説明図。

【図７】合成された音声波形の説明図。

【図８】音声モーフィング処理の概要説明図。

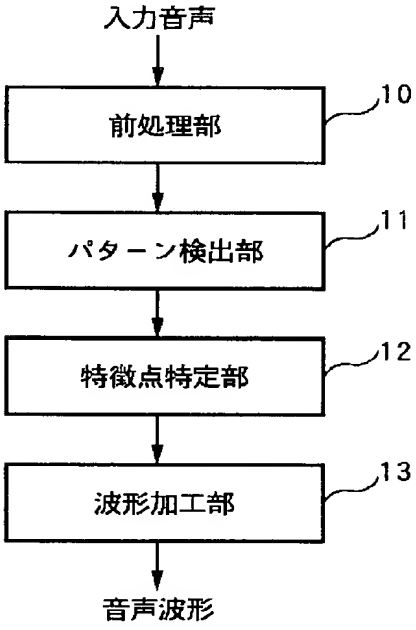
【図９】従来例における音声合成処理の手順説明図。

【図１０】（ａ），（ｂ）は人物Ａ，Ｂの音声についての音韻のピッチパターンを示すグラフ、（ｃ）は合成音についての音韻のピッチパターンを示すグラフ。

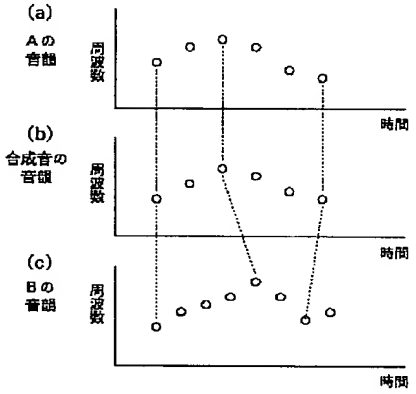
【符号の説明】

１０　前処理部
１１　パターン検出部
１２　特徴点特定部
１３　波形加工部

【図１】

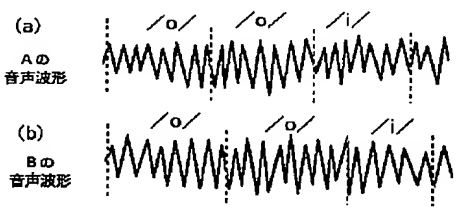


【図２】

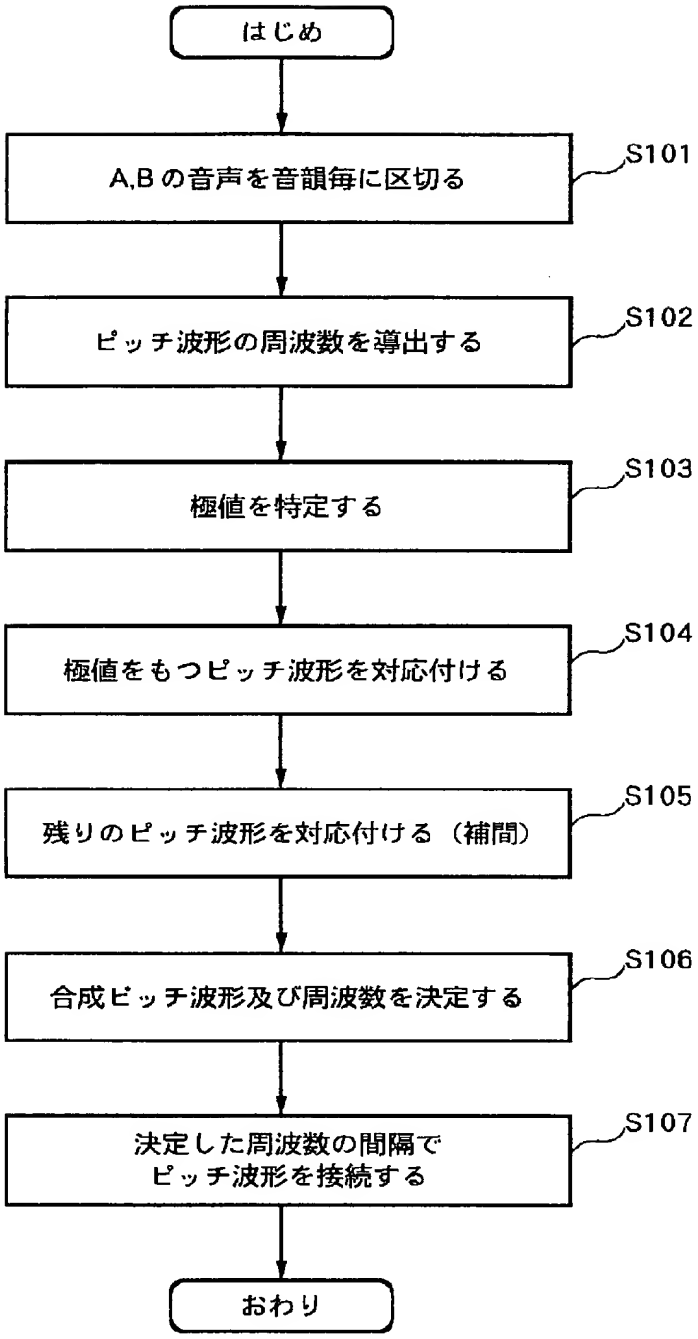


【図７】

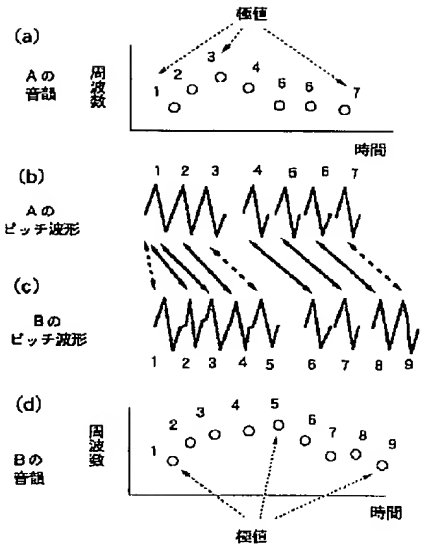
【図４】



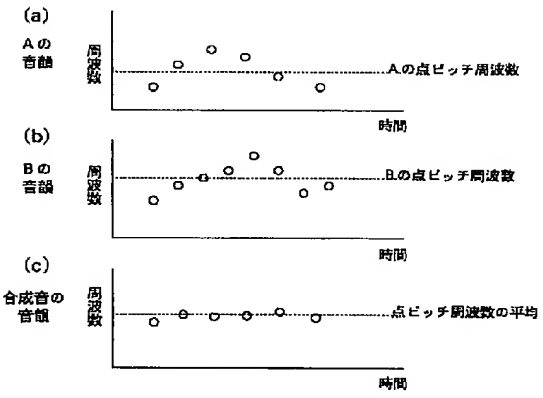
【図３】


はじめ
A,Bの音声を音韻毎に区切る
S101
ピッチ波形の周波数を導出する
S102
極値を特定する
S103
極値をもつピッチ波形を対応付ける
S104
残りのピッチ波形を対応付ける（補間）
S105
合成ピッチ波形及び周波数を決定する
S106
決定した周波数の間隔で
ピッチ波形を接続する
S107
おわり

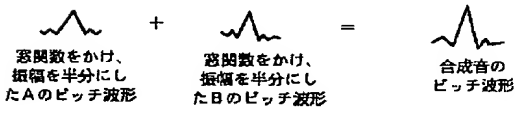
【図５】


極値
(a)
Aの音韻
周波数
時間
(b)
Aのピッチ波形
(c)
Bのピッチ波形
(d)
Bの音韻
周波数
時間
極値

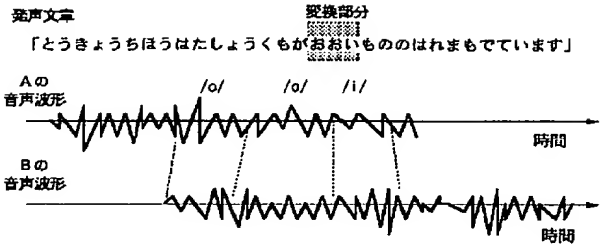
【図１０】


(a)
Aの音韻
周波数
Aの点ピッチ周波数
時間
(b)
Bの音韻
周波数
Bの点ピッチ周波数
時間
(c)
合成音の音韻
周波数
点ピッチ周波数の平均
時間

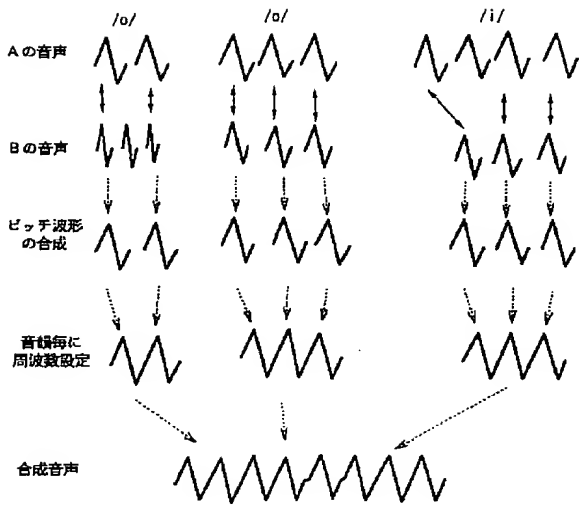
【図６】


窓関数をかけ、
振幅を半分にし
たAのピッチ波形
窓関数をかけ、
振幅を半分にし
たBのピッチ波形
合成音の
ピッチ波形

【図8】


発声文章
変換部分
「とうきょうちほうはたしょうくもがおおいもののはれまもでています」
Aの
音声波形
/o/
/o/
/i/
時間
Bの
音声波形
時間

【図9】


/o/
/o/
/i/
Aの音声
Bの音声
ピッチ波形
の合成
音韻毎に
周波数設定
合成音声